2016

Web Scraping the Easy Way

Yolande Neil

Web Scraping the Easy Way

An Honors Thesis submitted in partial fulfillment of the requirements for Honors in *Information Systems*

By
Yolande Neil

Under the mentorship of
Dr. John N. Dyer

ABSTRACT

Web scraping refers to a software program that mimics human web surfing behavior by pointing to a website and collecting large amounts of data that would otherwise be difficult for a human to extract. A typical program will extract both unstructured and semi-structured data, as well as images, and convert the data into a structured format. Web scraping is commonly used to facilitate online price comparisons, aggregate contact information, extract online product catalog data, extract economic/demographic/statistical data, and create web mashups, among other uses. Additionally, in the era of big data, semantic analysis, and business intelligence, web scraping is the only option for data extraction as many individuals and organizations need to consume large amounts of data that reside on the web. Although many users and organizations program their own web scrapers, there are scores of freely available programs and web-browser add-ins that can facilitate web scraping. This paper demonstrates web scraping using a free program named Data Toolbar® to extract data from Amazon.com. It is hoped that the paper will expose academicians, students and practitioners to not only the concept and necessity of web scraping, but the available software as well.

Thesis Mentor:_____

Dr. John N. Dyer

Honors Director:_____

Dr. Steven Engel

April 2016
Department of Information Systems
University Honors Program
Georgia Southern University

ACKNOWLEDGEMENTS

I would like to thank the University Honors Program at Georgia Southern University for giving me the opportunity to get involved with undergraduate research. Being able to have a platform to present my findings to is a great way for me to gain experience in how to undergo a complete research paper. My mentor, Dr. Dyer has played an essential role in coming up with the idea of utilizing a web scraping tool as the basis of this thesis. My parents have also been a great source of support with my journey into the Honors Program and the completion of this paper. Finally, I would like to thank DataTool Services for providing a free version of the Data Toolbar® software.

INTRODUCTION

There are many reasons why people and organizations want to scrape websites, and there are numerous web scraping programs available today. Organizations often seek web-based information that increases business value, including harnessing sales leads, market intelligence, news, creative content, company and sector performance data, enhanced e-commerce operations, and information for use in marketing and promotional campaigns.

In the modern era of big data and the need for data and information, people and companies alike are going to great lengths to gather relevant data and information. For example, Shopzilla® operates a portfolio of shopping websites that aggregate product availability and provides price comparisons for retail consumers; a half-billion-dollar company built on web scraping. Other companies like Nextag® and PriceGrabber® provide similar services.

A quick Internet search will yield numerous web scraping tools, from free and paid desktop applications to web-browser add-ins. While many websites provide an application program interface (API) or web-services to provide data to the client, many simply do not. Even when the API or web-service is provided, many individuals and smaller organizations do not have the technology and/or programming skill resources that are available in larger organizations. For example, both an API and a web-service require the client to write their own program according to the server's protocols and specifications. In the event the client doesn't have the capacity to consume the services then web scraping may be the only alternative. And while one may program their own

web scraper, there is really no need to since so many programs are already available. In this case the user simply adopts a web scraper, shows the web scraper how to navigate the website and the data to extract, and then the web scraper does the rest.

This paper is not novel contribution to the literature. Instead, it was motivated by the actual need of the author's client, and since web scraping is becoming a major tool on the landscape of data extraction it is important to disseminate the information. The aforementioned client is an online retailer aggregating products from numerous vendors for online resale. As an aggregator, the client simply advertises available vendor products in an online catalog and auction venues. The client does not hold inventory, but instead pre-sells available inventory and purchases the inventory in a just-in-time fashion. Some products are received by the client and shipped to customers, while other products are drop-shipped directly to the customer by the vendor. Unfortunately, few of the vendors made their inventory data available to retailers through an API or web-service, and only provided a dealer webpage to purchase products.

The client requested that the author create a web scraping program to aggregate as much of the product data as possible from as many of the vendor's websites as possible. After many months and hundreds of hours of programming, web scraping programs were developed for only two of the vendors. The task of writing programs for the remaining vendors simply wasn't feasible. By chance, the paper's co-author had an opportunity to see the programs being written and asked the author why so much e was being put into programming when commercial web scraping software was available at low and no cost. After investigating web scraper software solutions that author had an "aha moment." What took hundreds of hours of programming was accomplished in a matter of only a

few hours. Hence the motivation for this paper; to expose academicians, students and practitioners to the availability of easy to use web scraping software to accomplish massive data extraction from the web. Again, not a novel contribution, but instead, hopefully a timely topic.

BUSINESS INTELLIGENCE AND ANALYTICS

In order to truly understand the impact that business intelligence has had on the academic and business communities, it is crucial to first define the associated field of big data. The term 'big data' is said to have been coined by John Mashey and others at Silicon Graphics during the production of related work in the mid-1990s. This expression is used to reference data sets that are so large and complex, that it is necessary to utilize advanced storage, management analysis and visualization technologies. This is where business intelligence and analytics (BI&A) is needed to resolve these issues.

Even now, companies are gathering more data than they know what to do with. Organizations that have not been able to successfully make use of this data rely on the intuition of top decision-makers. Decision based on data are overall better as it is ingrained in evidence. If more managers were able to make well thought out decisions based on data, then all of this information would give them a competitive advantage. The utilization of web scraping tools could be the answer that many of these organizations are looking for.

WEB SCRAPING SOFTWARE

Web scraping is a software program that extracts information form websites. These programs are able to simulate human web surfing behavior by implementing either low-level Hypertext Transfer Protocol (HTTP) or embedding a web browser. Web scraping concentrates on transforming unstructured data found online, into structured data that can then be stored and analyzed in a database or spreadsheet program. This technique proves to be a welcomed change from the process of manually trying to gather large amounts of data.

Web data extraction software usually execute five different functions. First, they undergo the task of web interaction. This entails the software's navigation to the web pages that contain the desired information. Secondly, programs known as wrappers need support for generation and execution. The wrapper generation process typical has a visual interface that enables the user to identify the data that needs to be extracted from web pages and how it will be transformed into a structured format. Wrapper execution entails the deployment of previously generated wrappers.

Scheduling is the third function that allows wrappers to be applied repeatedly to their respective target pages. Next, data transformation occurs which involves the filtering, mapping, refining, and integrating of data from one or more sources. The result is then structured to create the desired output format. Finally, web extraction software should be able to deliver the resulting structured data to external application such as database management systems, data warehouses, business software systems, content management systems, decision support systems, RSS publishers, email servers, or SMS servers.

Alternatively, the output can be used to generate new web services out of existing and continually changing web sources.

CASE STUDY

The web scraping software used as the base model for this research was the Data Toolbar®. In this demonstration I will be extracting data from Amazon.com. This software is readily available for use for the average computer user. This paper will highlight the free edition of version 3.1 released in 2016 which allows you to run data scraping in the background, so that it does not interrupt other applications. The program does not require registration to use and is ad-free. It offers the same functionality as the full edition expect that the output is limited to 100 rows.

Once the program is up and running the default browser on your computer will appear with a DataTool button on the top left of the screen, as shown in figure 1. However, you can choose from either Internet Explorer, Mozilla Firefox or Google Chrome. The creation of a new project entails navigating to target web page and selecting the DataTool button. This step opens the DataTool window, which is known as the project designer. Once it is open you will be able to add content freely to the pre-defined template. Your mouse and keyboard will act as emulators and highlight the various page elements you will need for detailing your web scraping options.

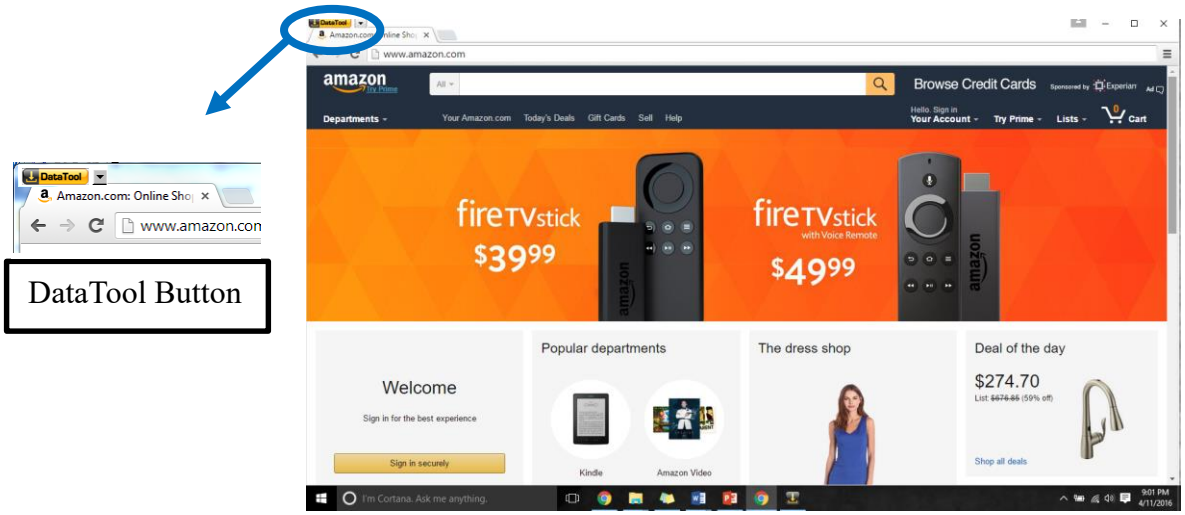


Figure 1. Initializing a web scraping project within the browser

Figure 2 below highlights the search groups and the criteria that was used in each. Field_01 represents the department drop-down where the term “books” was entered. Field_02 represents the search bar field where “paperback children’s books” was searched for. Finally, Field_03 represents the search button “Go.” When the “Open” action associated with this button is pressed, the program will start a search request on Amazon.com using the data values. Now you will be able to view the search results on the browser.

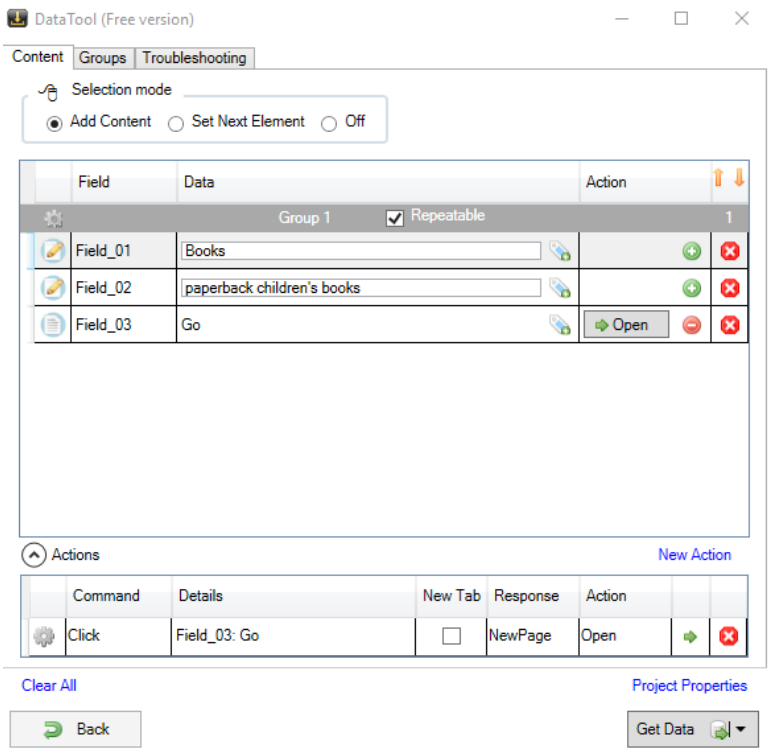


Figure 2. Creating the search form

Now the DataTool window will open an empty template where you add the page elements needed to gather data on. Figure 3 displays the elements where data will be collected from and titles were added to make the output look more organized. In order to ensure that the software would collect data from numerous pages, you must change the Selection Mode to “Set Next Element.” Field_04 represents the Next Page link that has the “Iterate” action associated with it. This ensures that the program will navigate to the other search results. Finally, the Get Data button is selected.

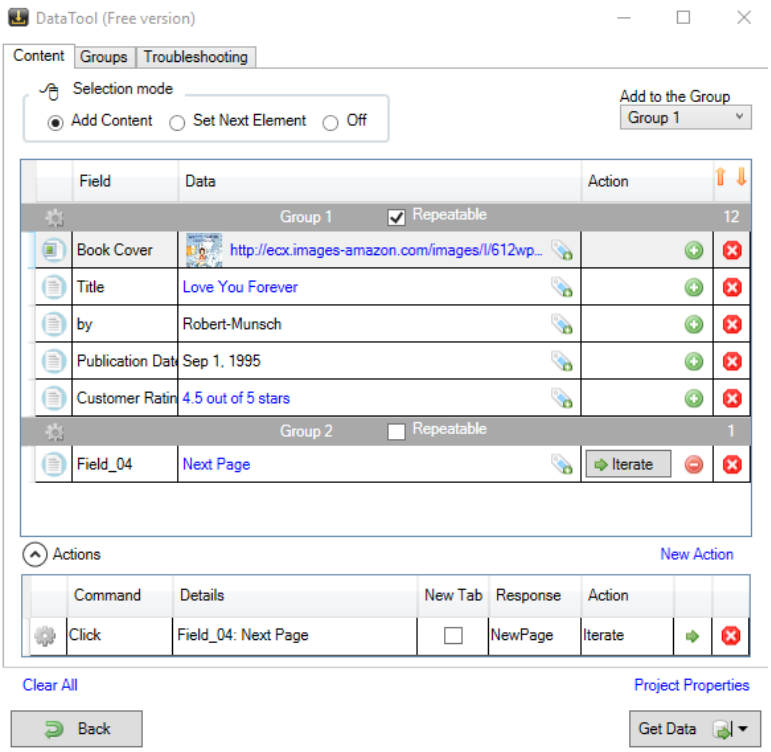


Figure 3. Defining input fields

Once the tool is finished processing all the pages, you will be able to review the collected data. The program will then, display in Review Data mode where each column is represented by each of the fields that were previously defined, while each book represents a single row of data. From here you can choose the location of your output files and choose the data type that you want it saved as. For the purposes of this paper, figure 4 displays what the information would look in an HTML format.

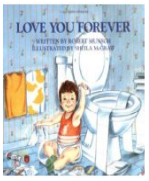
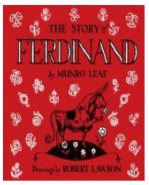
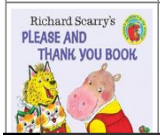
Book Cover	Title	by	Publication Date	Customer Rating
	Love You Forever	Robert-Munsch	Sep 1, 1995	4.5 out of 5 stars
	The Story of Ferdinand	Munro-Leaf	Mar 31, 2011	4.7 out of 5 stars
	May I Please Have a Cookie? (Scholastic Readers, Level 1)		Oct 1, 2005	4.8 out of 5 stars
	Richard Scarry's Please and Thank You Book (Pictureback(R))	Richard-Scarry	Aug 12, 1973	4.4 out of 5 stars

Figure 4. HTML output

CONCLUSION

The Data Toolbar® program is an intuitive web scraping tool that automates web data extraction process for your browser. It is designed for everyday business users and requires no technical skill. The program was able to gather most of the data that I had defined. That being said, I was not able to extract author names for every book. This was due to the fact that I had only selected the hyperlink element that displayed the name of a book's author. After taking another look at the list of search results I realized that some of the hyperlinks that contained the book author were not active.

FUTURE WORK

In the realm of future work, I hope to continue research of web scraping with more advanced software. This would allow me to use a more precise method of parsing that would ensure that every piece of data I wanted was collected. The ideal tool would allow me to create a unique program to extract the information that I needed. Overall, it is hoped that this body of work was able to expose academicians, students and practitioners to the concept and necessity of web scraping and demonstrate a tool that the average computer user could utilize on their own.

BIBLIOGRAPHY

Banerjee, Ritu. "Website Scraping." Happiest Minds. N.p., Apr. 2014. Web. 11 Apr. 2016.

Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS Quarterly 36.4 (2012): 1165-188. Web. 8 Apr. 2016.

Data Toolbar. Computer software. Data Toolbar. Vers. 3.1. DataTool Services Inc, 2013. Web. 8 Apr. 2016.

Diebold, Francis X., On the Origin(s) and Development of the Term 'Big Data' (September 21, 2012). PIER Working Paper No. 12-037. Web. 13 Apr. 2016.

Huynh, David, Stefano Mazzocchi, and David Karger. "Piggy Bank: Experience the Semantic Web Inside Your Web Browser." The Semantic Web – ISWC 2005 Lecture Notes in Computer Science (2005): 413-30. Web. 13 Apr. 2016.

Laender, Alberto H. F., Berthier A. Ribeiro-Neto, Altigran S. Da Silva, and Juliana S. Teixeira. "A Brief Survey of Web Data Extraction Tools." ACM SIGMOD Record SIGMOD Rec. 31.2 (2002): 84. Web. 11 Apr. 2016.

McAfee, Andrew, and Erik Brynjolfsson. "Big Data: The Management Revolution." Harvard Business Review. Hank Boye, 01 Oct. 2012. Web. 08 Apr. 2016.

Vargiu, Eloisa, and Mirko Urru. "Exploiting Web Scraping in a Collaborative Filtering- Based Approach to Web Advertising." Artificial Intelligence Research AIR 2.1 (2012): 44-54. Web. 13 Apr. 2016.